Discussion of “Post-Enumeration Data Processing Status”: Jay Breidt, Colorado State University

- Processing flow and key milestones: overall reaction is relief!
- Post-processing takes time
  - Every Census includes new methodologies not yet tested at full scale
  - Every complex operation encounters some unanticipated scenarios
  - Bugs happen even in well-developed software [aside: there exist statistical/machine learning/AI techniques to sniff out bugs in code]
- At the Fall 2020 meeting, CSAC…
  - Concluded that the compressed post-processing timeline proposed at that time (on top of pandemic conditions, natural disasters and other disruptions, and late process changes) threatened the accuracy of the 2020 Census
  - Recommended “…*that the Census Bureau have the time [the full six months] it requested in April 2020 to execute its full battery of data checks to reduce the risk of failing to identify key errors and generate final 2020 Census products that are of comparable quality to previous decennial censuses.*”
- Current processing timeline is consistent with Fall 2020 CSAC recommendation

External review and communications strategy

- JASON (independent group of scientists and engineers) report: JSR-20-2N (WS'21) Assessment of 2020 Census Data Quality Processes
- American Statistical Association Quality Indicators Task Force
  - plus special access for ASA members Paul Biemer, Robert Fay, and Joseph Salvo, each with Special Sworn Status
- Possible engagement with National Academy of Sciences (NAS) Committee on National Statistics
- Regular blogs on data processing and dissemination of external reviews (e.g., Jarmin 11/05 & 02/02; Thieme 02/11 & 03/09; Stempkowski 02/23; etc.)

Lessons learned?

- Great external review and efforts toward transparency
  - Keep up the good work!
  - Ensure that future iterations are planned, not reactive in appearance or in fact
- Continuously update test cases for future processing
  - *Standard problems that arise in processing any large survey [age error is a little surprising]*
  - *Anomalies resulting from unanticipated respondent action*
  - *Anomalies resulting from unanticipated enumerator action*
- Anomalytown, USA:
  - Artificial suite of test cases with lots of problems (real and simulated errors, edge cases, corner cases, duplications, disruptions, …) and desired outputs
  - Use in pre-testing and in regression testing as bugs are addressed

Defining “usually”?

- Does this person usually live or stay somewhere else?
  - *Here, you marked all that apply: no; yes, for college; yes, for a military assignment; yes, for a job or business; yes, in a nursing home; yes, with a parent or other relative; yes, at a seasonal or second residence; yes, in a jail or prison; yes, for another reason.*
  - *[https://2020census.gov/en/about-questions.html]*
- **Usually**: under normal conditions; generally.
- Better guidance on the temporal window that defines “usually”?
- More examples of disruptions, displacements?
- Might help with some unanticipated actions by respondent or enumerator